PCT/IB2004/051903

PHUS030388WO

CLAIMS:

1. An apparatus in a digital communication system that is capable of receiving audio-visual input signals that represent a speaker who is speaking and capable of creating an animated version of the face of the speaker using a plurality of audio logical units that represent the speaker's speech, said apparatus comprising a content synthesis application processor that:

extracts audio features of the speaker's speech and visual features of the speaker's face from the audio-visual input signals;

creates audiovisual input vectors from the audio features and the visual features;

creates audiovisual configurations from the audiovisual input vectors; and

performs a semantic association procedure on the audiovisual input vectors to obtain an association between phonemes that represent the speaker' speech and visemes that represent the speaker's face.

2. An apparatus as claimed in Claim 1 wherein the content synthesis application processor is capable of analyzing an input audio signal by:

extracting audio features of a speaker's speech;

finding corresponding video representations for the audio features using a semantic association procedure; and

matching the corresponding video representations with the audiovisual configurations.

3. An apparatus as claimed in Claim 2 wherein the content synthesis application processor is further capable of:

creating a computer generated animated face for each selected audiovisual configuration;

synchronizing each computer generated animated face with the speaker's speech; and

outputting an audio-visual representation of the speaker's face synchronized with the speaker's speech.

4. An apparatus as claimed in Claim 1 wherein the audio features that the content synthesis application processor extracts from the audio-visual input signals comprise one of: Mel Cepstral Frequency Coefficients, Linear Predictive Coding Coefficients, Delta Mel Cepstral Frequency Coefficients, Delta Linear Predictive Coding Coefficients, and Autocorrelation Mel Cepstral Frequency Coefficients.

5. An apparatus as claimed in Claim 1 wherein said content synthesis application processor creates audiovisual configurations from the audiovisual input vectors using one of: a Hidden Markov Model and a Time Delayed Neural Network.

6. An apparatus as claimed in Claim 2 wherein said content synthesis application processor matches the corresponding video representations with the audiovisual configurations using one of: a Hidden Markov Model and a Time Delayed Neural Network.

7. An apparatus as claimed in Claim 3 wherein said content synthesis application processor further comprises:

a facial audio visual feature matching and classification module that matches each of a plurality of audiovisual configurations with a corresponding classified audio feature to create a facial animation parameter; and

a facial animation for selected parameters module that creates an animated version of the face of the speaker for a selected facial animation parameter.

8. An apparatus as claimed in Claim 7 wherein said facial animation for selected parameters module creates an animated version of the face of the speaker by using one of: (1) 3D models with texture mapping and (2) video editing.

9. An apparatus as claimed in Claim 2 wherein said semantic association procedure comprises one of: latent semantic indexing, canonical correlation, and cross modal factor analysis.

PCT/IB2004/051903

10. An apparatus as claimed in Claim 1 wherein said audiovisual configurations comprise audiovisual speaking face movement components.

11. An apparatus as claimed in Claim 8 wherein said content synthesis application processor further comprises:

a speaking face animation and synchronization module that synchronizes each animated version of the face of the speaker with the audio features of the speaker's speech to create an audio-visual representation of the speaker's face that is synchronized with the speaker's speech; and

an audio expression classification module that determines a level of audio expression of the speaker's speech and provides said level of audio expression of the speaker's speech to said speaking face animation and synchronization module to use to modify animated facial parameters of the speaker.

12. A method for use in synthesizing audio-visual content in a video image processor, said method comprising the steps of:

receiving audio-visual input signals that represent a speaker who is speaking;

extracting audio features of the speaker's speech and visual features of the speaker's face from the audio-input signals;

creating audiovisual input vectors from the audio features and the visual features;

creating audiovisual configurations from the audiovisual input vectors; and

performing a semantic association procedure on the audiovisual input vectors to obtain an association between phonemes that represent the speaker' speech and visemes that represent the speaker's face.

13. The method as claimed in Claim 12 further comprising the steps of:

analyzing an input audio signal of a speaker's speech:

extracting audio features of the speaker's speech;

finding corresponding video representations for the audio features using a semantic association procedure; and

matching the corresponding video representations with the audiovisual configurations.

14. The method as claimed in Claim 13 further comprising the steps of:

creating a computer generated animated face for each selected audiovisual configuration;

synchronizing each computer generated animated face with the speaker's speech; and

outputting an audio-visual representation of the speaker's face synchronized with the speaker's speech.

15. The method as claimed in Claim 12 wherein the audio features that are extracted from the audio-visual input signals comprise one of: Mel Cepstral Frequency Coefficients, Linear Predictive Coding Coefficients, Delta Mel Cepstral Frequency Coefficients, Delta Linear Predictive Coding Coefficients, and Autocorrelation Mel Cepstral Frequency Coefficients.

16. The method as claimed in Claim 12 wherein the audiovisual configurations are created from the audiovisual input vectors using one of: a Hidden Markov Model and a Time Delayed Neural Network.

17. The method as claimed in Claim 13 wherein the corresponding video representations are matched with the audiovisual configurations using one of: a Hidden Markov Model and a Time Delayed Neural Network.

18. The method as claimed in Claim 12 further comprising the steps of:

matching each of a plurality of audiovisual configurations with a corresponding classified audio feature to create a facial animation parameter; and

creating an animated version of the face of the speaker for a selected facial animation parameter.

19. The method as claimed in 18 further comprising the step of:

creating an animated version of the face of the speaker by using one of: (1) 3D models with texture mapping and (2) video editing.

PCT/IB2004/051903

20. The method as claimed in Claim 13 wherein said semantic association procedure comprises one of: latent semantic indexing, canonical correlation, and cross modal factor analysis.

21. The method as claimed in Claim 12 wherein said audiovisual configurations comprise audiovisual speaking face movement components.

22. The method as claimed in Claim 20 further comprising the steps of:

synchronizing each animated version of the face of the speaker with the audio features of the speaker's speech;

creating an audio-visual representation of the face of the speaker that is synchronized with the speaker's speech;

determining a level of audio expression of the speaker's speech; and

modifying animated facial parameters of the speaker in response to a determination of the level of audio expression of the speaker's speech.

23. A synthesized audio-visual signal generated by a method for synthesizing audio-visual content in a video image processor, wherein the method comprises the steps of:

receiving audio-visual input signals that represent a speaker who is speaking;

extracting audio features of the speaker's speech and visual features of the speaker's face from the audio-input signals;

creating audiovisual input vectors from the audio features and the visual features;

creating audiovisual configurations from the audiovisual input vectors; and

performing a semantic association procedure on the audiovisual input vectors to obtain an association between phonemes that represent the speaker' speech and visemes that represent the speaker's face.

24. A synthesized audio-visual signal as claimed in Claim 23 wherein the method further comprises the steps of:

analyzing an input audio signal of a speaker's speech:

extracting audio features of the speaker's speech;

finding corresponding video representations for the audio features using a semantic association procedure; and

matching the corresponding video representations with the audiovisual configurations.

25. A synthesized audio-visual signal as claimed in Claim 24 wherein the method further comprises the steps of:

creating a computer generated animated face for each selected audiovisual configuration;

synchronizing each computer generated animated face with the speaker's speech; and

outputting an audio-visual representation of the speaker's face synchronized with the speaker's speech.

26. A synthesized audio-visual signal as claimed in Claim 23 wherein the audio features that are extracted from the audio-visual input signals comprise one of: Mel Cepstral Frequency Coefficients, Linear Predictive Coding Coefficients, Delta Mel Cepstral Frequency Coefficients, Delta Linear Predictive Coding Coefficients, and Autocorrelation Mel Cepstral Frequency Coefficients.

27. A synthesized audio-visual signal as claimed in Claim 23 wherein the audiovisual configurations are created from the audiovisual input vectors using one of: a Hidden Markov Model and a Time Delayed Neural Network.

28. A synthesized audio-visual signal as claimed in Claim 24 wherein the corresponding video representations are matched with the audiovisual configurations using one of: a Hidden Markov Model and a Time Delayed Neural Network.

29. A synthesized audio-visual signal as claimed in Claim 25 wherein the method further comprises the steps of:

matching each of a plurality of audiovisual configurations with a corresponding classified audio feature to create a facial animation parameter; and

creating an animated version of the face of the speaker for a selected facial animation parameter.

PCT/IB2004/051903

30. A synthesized audio-visual signal as claimed in Claim 29 said method further comprises the step of:

creating an animated version of the face of the speaker by using one of: (1) 3D models with texture mapping and (2) video editing.

31. A synthesized audio-visual signal as claimed in Claim 24 wherein said semantic procedure comprises one of: latent semantic indexing, canonical correlation, and cross modal factor analysis.

32. A synthesized audio-visual signal as claimed in Claim 23 wherein said audiovisual configurations comprise audiovisual speaking face movement components.

33. A synthesized audio-visual signal as claimed in Claim 31 wherein the method further comprises the steps of:

synchronizing each animated version of the face of the speaker with the audio features of the speaker's speech;

creating an audio-visual representation of the face of the speaker that is synchronized with the speaker's speech

determining a level of audio expression of the speaker's speech; and

modifying animated facial parameters of the speaker in response to a determination of the level of audio expression of the speaker's speech.